

ΕΜΠ ΔΠΜΣ

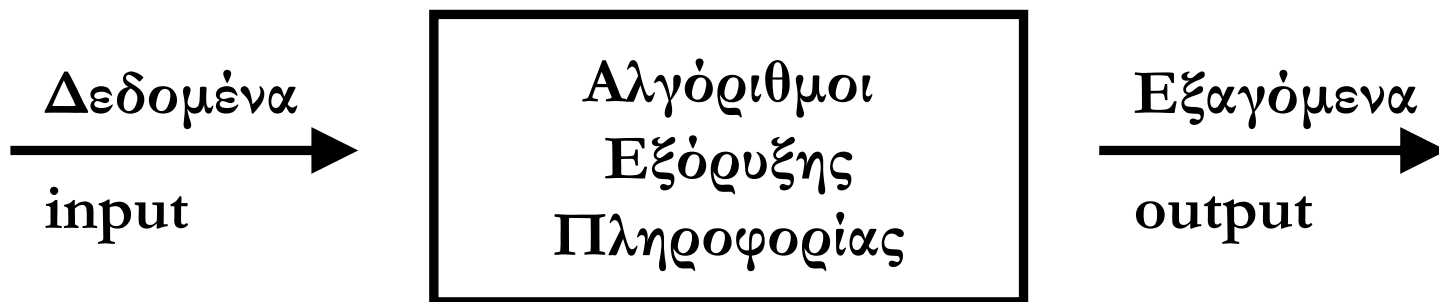
Εφαρμοσμένες Μαθηματικές Επιστήμες
Αλγόριθμοι Εξόρυξης Πληροφορίας

Διάλεξη 02

Συνιστώσες Δεδομένων
Οπτικοποίηση & Εξερεύνηση



Η μορφή των δεδομένων και η σημασία της



- Η κατανόηση της μορφής των δεδομένων και εξαγόμενων είναι ίσως περισσότερο σημαντική από τα ενδιάμεσα βήματα σε έναν αλγόριθμο εξόρυξης πληροφορίας
- Αντικείμενο εστίασης των πρώτων διαλέξεων



Συνιστώσες Δεδομένων Ορολογία

- **Αντίληψη (concept):** το αντικείμενο της μάθησης
 - Στόχος: εύρεση κατανοητής και λειτουργικής περιγραφής ενός concept
- **Υπόδειγμα (instance):** το ξεχωριστό και ανεξάρτητο παράδειγμα (example) ενός concept
 - Σημείωση: Περισσότερο σύνθετα σχήματα εισόδου είναι πιθανά
- **Χαρακτηριστικό (attribute):** η μετρήσιμη συνιστώσα ενός υποδείγματος
 - Θα εστιάσουμε σε ονομαστικές και αριθμητικές συνιστώσες



Συνιστώσες Δεδομένων

Παράδειγμα

← attributes →

	Outlook	Temperature	Humidity	Windy	Play
↑ instances ↓	Sunny	85	85	False	No
	Sunny	80	90	True	No
	Overcast	83	86	False	Yes
	Rainy	75	80	False	Yes

concept description:

If outlook = sunny and humidity = high then play = no



Concept (Αντίληψη)

- Είδη μάθησης:
 - Ταξινόμηση:
πρόβλεψη διακριτής κατηγορίας
 - Συσχέτιση:
εντοπισμός συσχετίσεων μεταξύ χαρακτηριστικών
 - Ομαδοποίηση:
ανάδειξη ομάδων όμοιων υποδειγμάτων
 - Αριθμητική πρόβλεψη:
πρόβλεψη αριθμητικής ποσότητας
- Concept (αντίληψη): το αντικείμενο της μάθησης
- Περιγραφή αντίληψης (concept description):
το προϊόν / σχήμα της μαθησιακής διαδικασίας



Ταξινόμηση (Classification)

- Ταξινόμηση υποδείγματος σε προκαθορισμένη τάξη (*class*)
- Παραδείγματα προβλημάτων: δεδομένα καιρού, φακοί επαφής, ίρις, εργασιακές διαπραγματεύσεις
- Μάθηση με επίβλεψη (*supervised*)
 - Οι ομάδες ταξινόμησης είναι εκ των προτέρων γνωστές
 - Το πραγματικό αποτέλεσμα κάθε υποδείγματος είναι επίσης γνωστό
- Ο βαθμός αξιοπιστίας μετριέται
 - σε μη χρησιμοποιημένα για τη διαμόρφωση του concept description δεδομένα (*test data*) είτε
 - Υποκειμενικά, ανάλογα με το βαθμό αποδοχής της περιγραφής



Συσχέτιση (Association)

- Ανακάλυψη συσχετίσεων μεταξύ των διάφορων χαρακτηριστικών
- Διαφορές με τη μάθηση ταξινόμησης:
 - Μπορεί να αναδείξουν τη συσχέτιση είτε να προβλέψουν την τιμή οποιουδήποτε χαρακτηριστικού και όχι μόνο της τάξης
 - Συνδέουν πιθανόν περισσότερα από ένα χαρακτηριστικά κάθε φορά
 - Επομένως, προικύπτουν πολλοί περισσότεροι κανόνες συσχέτισης από ότι κανόνες ταξινόμησης
 - Άρα περιορισμοί κατά την αναζήτηση είναι αναγκαίοι, όπως ελάχιστη κάλυψη & ελάχιστη ακρίβεια



Ομαδοποίηση (Clustering)

- Εύρεση ομάδων αντικειμένων με υψηλό βαθμό ομοιότητας και ειχώρηση υποδειγμάτων στις ομάδες αυτές
- Χωρίς επίβλεψη (*unsupervised*)
 - η τάξη του υποδείγματος δεν ανήκει σε γνωστό σύνολο

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
...					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
...					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
...					



Αριθμητική πρόβλεψη (Numeric prediction)

- Όμοια με ταξινόμηση, αλλά τώρα η 'τάξη' είναι αριθμητική
- Μάθηση με επίβλεψη (*supervised*)
 - Η τιμή – στόχος κάθε υποδείγματος είναι εκ των προτέρων γνωστή

Outlook	Temperature	Humidity	Windy	Play-time
Sunny	Hot	High	False	5
Sunny	Hot	High	True	0
Overcast	Hot	High	False	55
Rainy	Mild	Normal	False	40
...

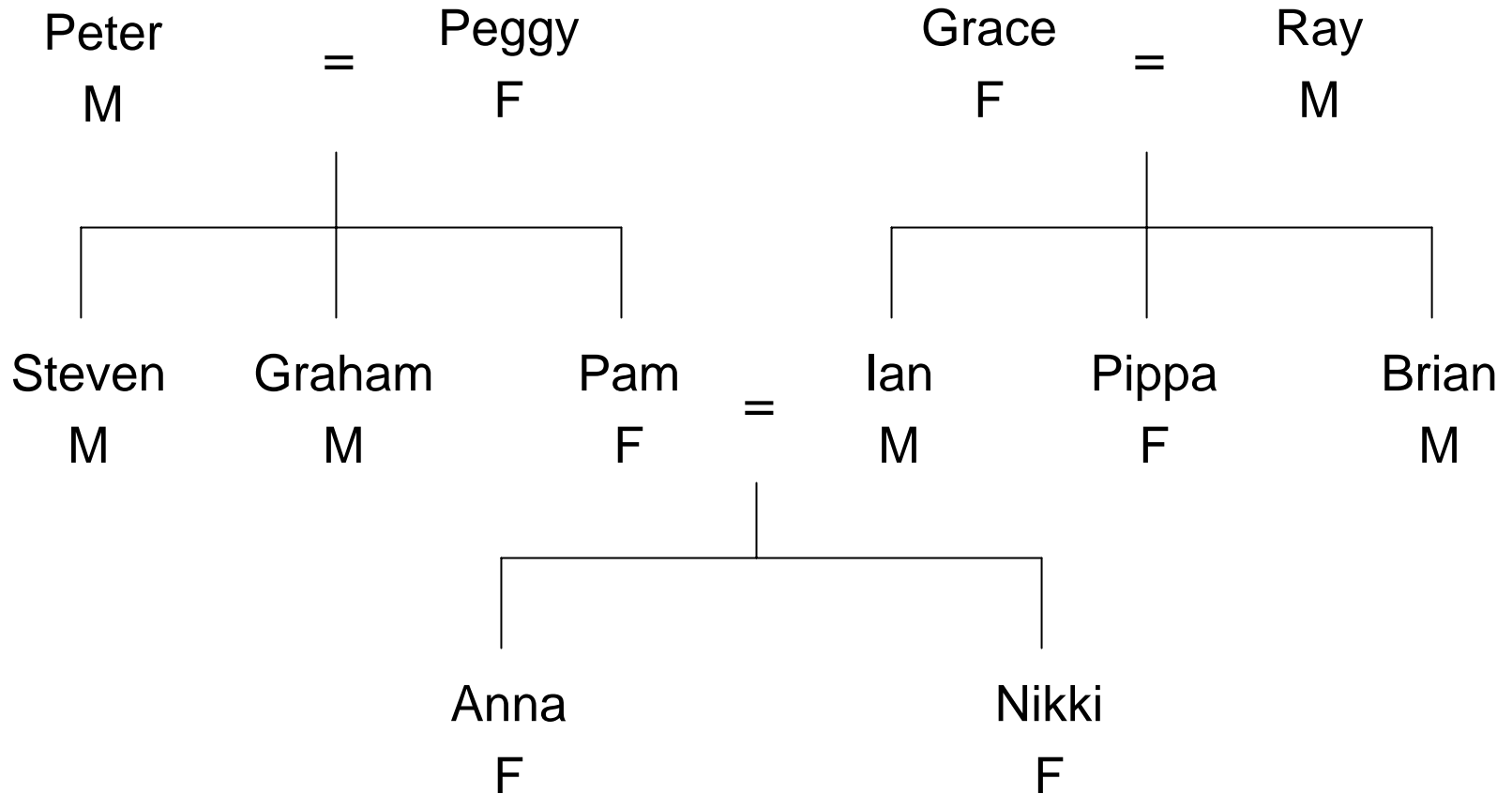


Παράδειγμα (example)

- Υπόδειγμα (instance): ειδικός τύπος παραδείγματος
 - Προς ταξινόμηση, συσχέτιση ή ομαδοποίηση
 - Ιδιαίτερο & ανεξάρτητο παράδειγμα της αντίληψης προς εκμάθηση
 - Χαρακτηρίζεται από προκαθορισμένο σύνολο χαρακτηριστικών
- Συνήθης μορφή δεδομένων προς εξόρυξη γνώσης: σύνολο υποδειγμάτων (dataset)
 - Πίνακας υποδειγμάτων (instances) – χαρακτηριστικών (attributes) (επίπεδο αρχείο, flat file)
 - Μάλλον περιοριστική μορφή δεδομένων (καμία συσχέτιση μεταξύ των αντικειμένων)



Οικογενειακό δένδρο






Οικογενειακό δένδρο σε μορφή πίνακα



Name	Gender	Parent1	Parent2
Peter	Male	?	?
Peggy	Female	?	?
Steven	Male	Peter	Peggy
Graham	Male	Peter	Peggy
Pam	Female	Peter	Peggy
Ian	Male	Grace	Ray
Pippa	Female	Grace	Ray
Brian	Male	Grace	Ray
Anna	Female	Pam	Ian
Nikki	Female	Pam	Ian



Μορφές απεικόνισης συσχέτισης

First person	Second person	Sister of?	First person	Second person	Sister of?
Peter	Peggy	No	Steven	Pam	Yes
Peter	Steven	No	Graham	Pam	Yes
...	Ian	Pippa	Yes
Steven	Peter	No	Brian	Pippa	Yes
Steven	Graham	No	Anna	Nikki	Yes
Steven	Pam	Yes	Nikki	Anna	Yes
...	<i>All the rest</i>		No
Ian	Pippa	Yes	 <i>Υπόθεση κλειστού κόσμου (Closed-world assumption)</i>		
...			
Anna	Nikki	Yes			
...			
Nikki	Anna	yes			



Μορφή πλήρους απεικόνισης σε πίνακα

First person				Second person				Sister of?
Name	Gender	Parent1	Parent2	Name	Gender	Parent1	Parent2	
Steven	Male	Peter	Peggy	Pam	Female	Peter	Peggy	Yes
Graham	Male	Peter	Peggy	Pam	Female	Peter	Peggy	Yes
Ian	Male	Grace	Ray	Pippa	Female	Grace	Ray	Yes
Brian	Male	Grace	Ray	Pippa	Female	Grace	Ray	Yes
Anna	Female	Pam	Ian	Nikki	Female	Pam	Ian	Yes
Nikki	Female	Pam	Ian	Anna	Female	Pam	Ian	Yes
<i>All the rest</i>								No

**If second person's gender = female
and first person's parent = second person's
parent
then sister-of = yes**



Δημιουργώντας ένα επίπεδο αρχείο

- Η διαδικασία καλείται ‘αποκανονικοποίηση’ (denormalization)
 - Για παράδειγμα, μετατροπή δένδρου σε πίνακα μέσω συγχώνευσης συσχετίσεων
- Εφικτή με κάθε πεπερασμένο σύνολο πεπερασμένων συσχετίσεων
- Πρόβλημα: Συσχετίσεις χωρίς προκαθορισμένο αριθμό αντικειμένων
 - Μία σειρά για κάθε συνδυασμό ατόμων
 - Υψηλό υπολογιστικό κόστος και κόστος αποθήκευσης
- Η αποκανονικοποίηση μπορεί να παράγει πλαστές ή προφανείς κανονικότητες που απεικονίζουν τη δομή της βάσης δεδομένων



Η συσχέτιση 'πρόγονος'

First person				Second person				Ancestor of?
Name	Gender	Parent1	Parent2	Name	Gender	Parent1	Parent2	
Peter	Male	?	?	Steven	Male	Peter	Peggy	Yes
Peter	Male	?	?	Pam	Female	Peter	Peggy	Yes
Peter	Male	?	?	Anna	Female	Pam	Ian	Yes
Peter	Male	?	?	Nikki	Female	Pam	Ian	Yes
Pam	Female	Peter	Peggy	Nikki	Female	Pam	Ian	Yes
Grace	Female	?	?	Ian	Male	Grace	Ray	Yes
Grace	Female	?	?	Nikki	Female	Pam	Ian	Yes
<i>Other positive examples here</i>								Yes
<i>All the rest</i>								No



Αναδρομή (recursion)

- Οι συσχετίσεις απείρου μήκους απαιτούν αναδρομή

**If person1 is a parent of person2
then person1 is an ancestor of person2**

**If person1 is a parent of person2
and person2 is an ancestor of person3
then person1 is an ancestor of person3**

- Κατάλληλες τεχνικές όπως επαγωγικός λογικός προγραμματισμός (inductive logic programming)
 - Προβλήματα: θόρυβος και υπολογιστική πολυπλοκότητα



Χαρακτηριστικό (attribute)

- Κάθε παράδειγμα αποτελείται από προκαθορισμένο σύνολο στοιχείων, τα οποία καλούνται χαρακτηριστικά (attributes)
- Στην πράξη, ο αριθμός των χαρακτηριστικών μπορεί να ποικίλει
- Επίσης, η ύπαρξη ενός χαρακτηριστικού μπορεί να εξαρτάται από την τιμή ενός άλλου
- Πιθανοί τύποι χαρακτηριστικών (“levels of measurement”):
 - Ονομαστικά (*nominal*), τακτικά (*ordinal*), περιοδικά (*interval*) και αναλογικά (*ratio*)



Ονομαστικά (nominal) χαρακτηριστικά

- Οι τιμές είναι διακριτά σύμβολα
 - Αποτελούν περιγραφή ή όνομα
- Παράδειγμα: χαρακτηριστικό “outlook” από το πρόβλημα καιρού
 - Τιμές: “sunny”, “overcast” & “rainy”
- Δεν υπάρχει ή συνεπάγεται καμία συσχέτιση (κατάταξη ή μέτρο απόστασης) μεταξύ των ονομαστικών τιμών
- Μόνο η σύγκριση ως προς την ισότητα της τιμής του χαρακτηριστικού στα διάφορα υποδείγματα έχει νόημα



Τακτικά (ordinal) χαρακτηριστικά

- Ορίζεται διάταξη στις τιμές
- Αλλά δεν ορίζεται απόσταση μεταξύ τους
- Για παράδειγμα, χαρακτηριστικό “temperature” στο πρόβλημα καιρού
 - Τιμές: “hot” > “mild” > “cool”
- Σημείωση: η πρόσθεση και η αφαίρεση δεν έχουν νόημα
- Παράδειγμα κανόνα:
temperature < hot → play = yes
- Η διάκριση μεταξύ ονομαστικών και τακτικών χαρακτηριστικών δεν είναι πάντα ευκρινής (για παράδειγμα “outlook”)



Περιοδικά (interval) χαρακτηριστικά

- Οι τιμές των περιοδικών χαρακτηριστικών δεν είναι μόνο διατεταγμένες, αλλά μετρώνται σε σταθερές και ισαπέχουσες μονάδες
- Παράδειγμα: χαρακτηριστικό “temperature” σε βαθμούς Fahrenheit
- Παράδειγμα: χαρακτηριστικό “έτος”
- Η διαφορά ανάμεσα σε δύο τιμές έχει έννοια
- Ωστόσο το άθροισμα ή το γινόμενο δεν έχει έννοια, καθώς δεν έχει οριστεί το σημείο μηδέν.



Αναλογικά (ratio) χαρακτηριστικά

- Η μέθοδος μέτρησης ορίζει σημείο μηδέν
- Παράδειγμα: χαρακτηριστικό “απόσταση”
 - Η απόσταση ανάμεσα σε ένα αντικείμενο και τον εαυτό του ορίζεται ως μηδενική
- Τα αναλογικά χαρακτηριστικά μεταχειρίζονται ως πραγματικοί αριθμοί
 - Όλες οι μαθηματικές πράξεις είναι επιτρεπτές
- Ωστόσο ο ορισμός του σημείου μηδέν είναι μάλλον σχετικός παρά απόλυτος
 - Για παράδειγμα, βαθμοί Fahrenheit



Στην πράξη

- Στις περισσότερες των περιπτώσεων, χρησιμοποιούνται δύο μόνο τύποι χαρακτηριστικών: ονομαστικά και τακτικά (nominal and ordinal)
- Τα ονομαστικά χαρακτηριστικά συχνά καλούνται επίσης ρητά (*categorical*), απαριθμημένα (*enumerated*) ή διακριτά (*discrete*)
 - Ωστόσο ο ακριβής ορισμός των ρητών και απαριθμημένων προϋποθέτει διάταξη
- Εξαιρέση: διχοτόμηση (δυναδικό (boolean) χαρακτηριστικό)
- Τα τακτικά χαρακτηριστικά καλούνται επίσης αριθμητικά (*numeric*) ή συνεχή (*continuous*)
 - Ωστόσο η συνέχεια παραπέμπει στον μαθηματικό ορισμό της



Μετασχηματισμός τακτικών σε δυαδικά χαρακτηριστικά

- Απλοί μετασχηματισμοί επιτρέπουν την κωδικοποίηση ενός τακτικού (ordinal) χαρακτηριστικού με n τιμές σε $n-1$ δυαδικά (boolean) χαρακτηριστικά
- Παράδειγμα: χαρακτηριστικό “temperature”

Δεδομένα (αρχικά)

Temperature
Cold
Medium
Hot



Δεδομένα μετά τον μετασχηματισμό

Temperature > cold	Temperature > medium
False	False
True	False
True	True

- Προτεινόμενη κωδικοποίηση έναντι ονομαστικού χαρακτηριστικού



Μεταδεδομένα (metadata)

- Πληροφορίες σχετικές με τα δεδομένα που κωδικοποιούν τη γνώση πεδίου (background knowledge)
- Μπορούν να χρησιμοποιηθούν για τον περιορισμό του χώρου αναζήτησης
- Παραδείγματα:
 - Θεώρηση διαστάσεων
(οι διαστάσεις των περιγραφών πρέπει να είναι συμβατές με εκείνες των δεδομένων)
 - Κυκλική διάταξη (circular ordering)
(για παράδειγμα, αριθμός μοιρών γωνίας)
 - Μερική διάταξη (partial ordering)
(για παράδειγμα, σχέσεις γενίκευσης ή εξειδίκευσης)



Προπαρασκευή δεδομένων εισόδου

- Πρόβλημα: διαφορετικές πηγές δεδομένων (για παράδειγμα, σε μία επιχείρηση, εγγραφές τμήματος πωλήσεων, λογιστηρίου, ...)
 - Διαφορές: τρόπος αποθήκευσης εγγραφών, παραδοχές, χρονικές περιοδοί, άθροιση δεδομένων, σφάλματα
 - Τα δεδομένα πρέπει να συγκεντρωθούν σε ενιαίο σύνολο, με ενιαία λιτή δομή
 - Αποθήκη δεδομένων (data warehouse): συνεπές σημείο πρόσβασης
- Εξωτερικά -προς την επιχείρηση- δεδομένα είναι συχνά αναγκαία (δεδομένα επικάλυψης / overlay data)
- Κρίσιμος παράγοντας: τύπος και επίπεδο άθροισης δεδομένων
 - Για παράδειγμα δεδομένα ανά συναλλαγή, ανά πελάτη, ανά ημέρα κτλ.



Άγνωστες τιμές

- Οι άγνωστες τιμές (missing values) συχνά υποδηλώνονται με καταχωρήσεις εκτός πεδίου τιμών (out-of-range)
 - Είδη: άγνωστες, μη καταγεγραμμένες, μη συσχετιζόμενες
 - Αίτια:
 - Βλάβη εξοπλισμού καταγραφής
 - Αλλαγές στο σχεδιασμό του πειράματος
 - Αντιπαράβολή διαφορετικών συνόλων δεδομένων
 - Μη εφικτή μέτρηση
- Συχνά χρειάζεται να κωδικοποιηθούν ως ξεχωριστή τιμή
- Κάποιες φορές έχουν ιδιαίτερη αξία από μόνες τους



Ανακριβείς τιμές

- Αίτιο: τα δεδομένα έχουν συλλεχθεί για σκοπό διάφορο της εξόρυξης γνώσης από αυτά
- Συνέπεια: λάθη & παραλείψεις που δεν επηρεάζουν τον αρχικό σκοπό των δεδομένων (για παράδειγμα ηλικία πελάτη)
- Τυπογραφικά λάθη σε ονομαστικά χαρακτηριστικά → απαιτείται έλεγχος αξιοπιστίας
- Τυπογραφικά λάθη & σφάλματα μέτρησης σε αριθμητικά χαρακτηριστικά → απαιτείται εντοπισμός των τιμών προς εξαίρεση (outliers)
- Τα λάθη μπορεί να είναι σιόπιμα (για παράδειγμα ψευδής Ταχυδρομικός Κώδικας)
- Άλλα προβλήματα: διπλότυπα, πεπαλαιωμένα δεδομένα



Εξερεύνηση των δεδομένων

- Ορισμένα απλά εργαλεία οπτικοποίησης είναι πολύ χρήσιμα
 - Ονομαστικά χαρακτηριστικά: ιστογράμματα (είναι η κατανομή συμβατή με τη γνώση πεδίου;)
 - Αριθμητικά χαρακτηριστικά: γραφήματα (υπάρχουν εμφανείς τιμές προς εξαίρεση (outliers);)
- Διαγράμματα 2 & 3 διαστάσεων υποδεικνύουν εξαρτήσεις & αλληλοσυσχετίσεις
- Αναγκαία η γνώση των ειδικών
- Αχανής όγκος δεδομένων; Δειγματοληψία!



Εισαγωγή στο weka



Λογισμικό - Γενικά

- Ο αλγόριθμος εξόρυξης καθολικής εφαρμογής αποτελεί ιδεαλιστική ουτοπία
- Τα σύνολα δεδομένων ποικίλουν ευρέως στην πράξη
- Για να καταλήξει κανείς σε αξιόπιστη μοντελοποίηση, απαιτείται συναρμογή των χαρακτηριστικών του αλγορίθμου εκμάθησης (όπως bias) με τη δομή του πεδίου εφαρμογής
- Η ανακάλυψη γνώσης από δεδομένα προϋποθέτει ειτενή εφαρμογή της μεθόδου trial & error και αποτελεί εξ' ορισμού πειραματική επιστήμη



Weka – το πτηνό





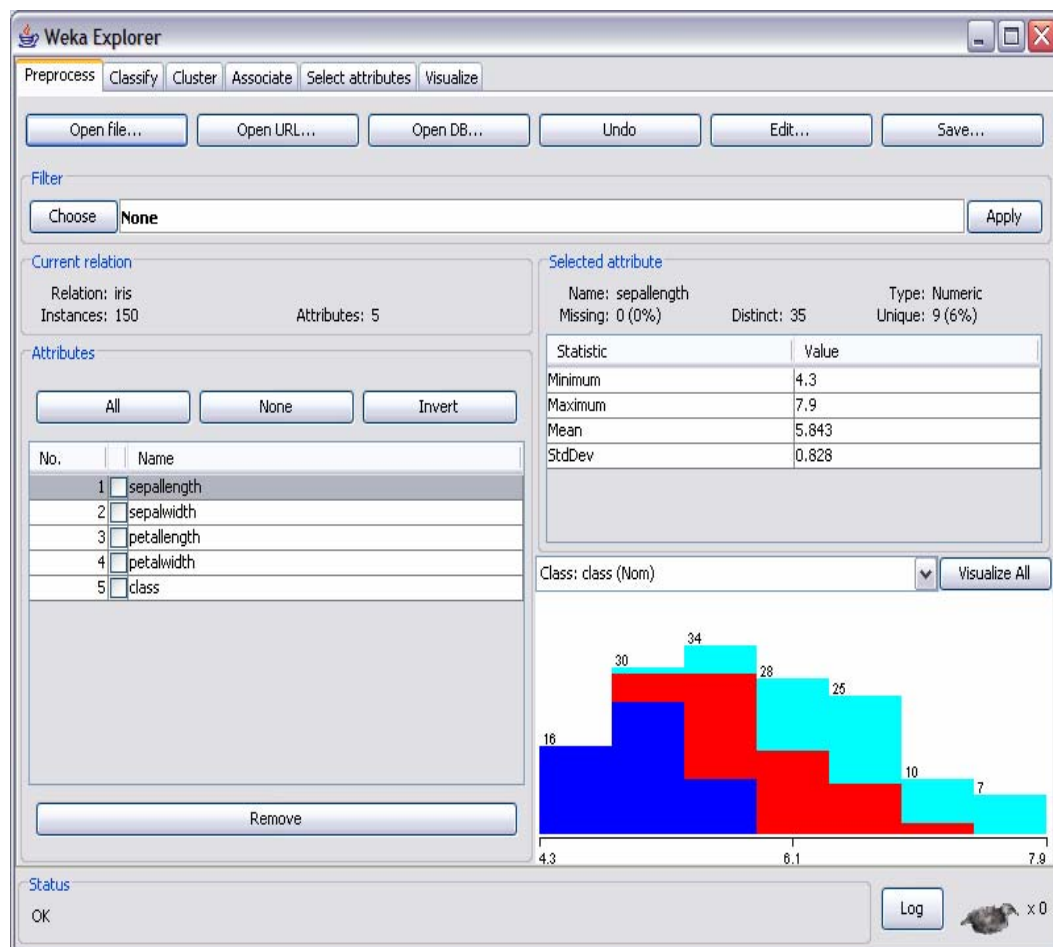
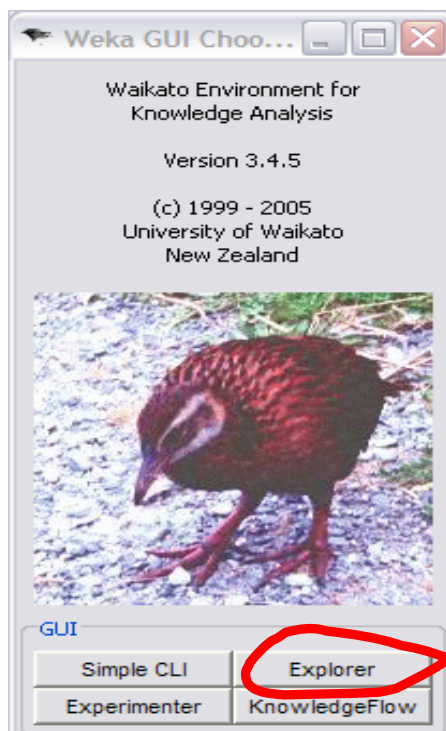
Weka – Το περιβάλλον εργασίας



- Waikato Environment for Knowledge Analysis
- open-source, γραμμένο σε Java
- Ειτενής συλλογή των πλέον σύγχρονων αλγορίθμων εξόρυξης πληροφορίας και εργαλείων προεπεξεργασίας δεδομένων
- Ενιαία και εύχρηστη διεπιφάνεια (workbench) για την υλοποίηση αυτών



GUI chooser & Explorer





Η τυποποίηση .ARFF



%

% ARFF file for weather data with some numeric features

%

@relation weather

@attribute outlook {sunny, overcast, rainy}

@attribute temperature numeric

@attribute humidity numeric

@attribute windy {true, false}

@attribute play? {yes, no}

@data

sunny, 85, 85, false, no

sunny, 80, 90, true, no

overcast, 83, 86, false, yes

...



Τύποι χαρακτηριστικών



- Η τυποποίηση .ARFF αποτελεί τη φυσική μέθοδο αποθήκευσης δεδομένων του weka
- Υποστηρίζει αριθμητικά και ονομαστικά χαρακτηριστικά
- Η ερμηνεία εξαρτάται από το είδος μάθησης
 - Τα αριθμητικά χαρακτηριστικά ερμηνεύονται συχνά ως
 - Τακτικά (ordinal) με συσχετίσεις μικρότερου / μεγαλύτερου
 - Αναλογικά (ratio) με συσχετίσεις απόστασης (ίσως απαιτεί κανονικοποίηση)
 - Τα ονομαστικά χαρακτηριστικά μπορούν να ερμηνευθούν με όρους απόστασης (0 εάν οι τιμές είναι ίσες, 1 διαφορετικά)
- Ανέρχονται αριθμοί: ονομαστική (nominal), τακτική (ordinal) ή αναλογική (ratio) κλίμακα?



Επιλογή τύπου χαρκτηριστικού

- Χαρκτηριστικό “age” ονομαστικού (nominal) τύπου
 - If age = young and astigmatic = no
and tear production rate = normal
then recommendation = soft
 - If age = pre-presbyopic and astigmatic =
no
and tear production rate = normal
then recommendation = soft
- Χαρκτηριστικό “age” τακτικού (ordinal) τύπου (για παράδειγμα “young” < “pre-presbyopic” < “presbyopic”)
 - If age \leq pre-presbyopic and astigmatic = no
and tear production rate = normal
then recommendation = soft



Μετατροπή .xls σε .arff



- Έστω πίνακας σε αρχείο .xls
- Save as → save as type → name.csv
- Open name.csv με .txt editor, για παράδειγμα notepad
- Πρόσθεσε το όνομα του dataset (*@relation*), τις πληροφορίες των χαρακτηριστικών (*@attribute*, μία σειρά για κάθε χαρακτηριστικό) και τη σειρά *@data*
- Save as type: all files & filename: dataset.arff



Εισαγωγή δεδομένων



- Το λογισμικό παρέχει τους εξής τρόπους εισαγωγής δεδομένων:
 - Αρχεία (.arff / .csv / .c45 / binary)
 - URL
 - Database (μέσω sql query)
- Στο φάκελο C:\Program Files\Weka-3-4\data είναι διαθέσιμα τα datasets όλων των παραδειγμάτων που έχουν αναφερθεί
- Ανοίξτε το dataset λουλουδιών ίρις (iris.arff)



Σύνολο δεδομένων λουλουδιών Ίρις



	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
...					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
...					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
...					



Εξερεύνηση δεδομένων



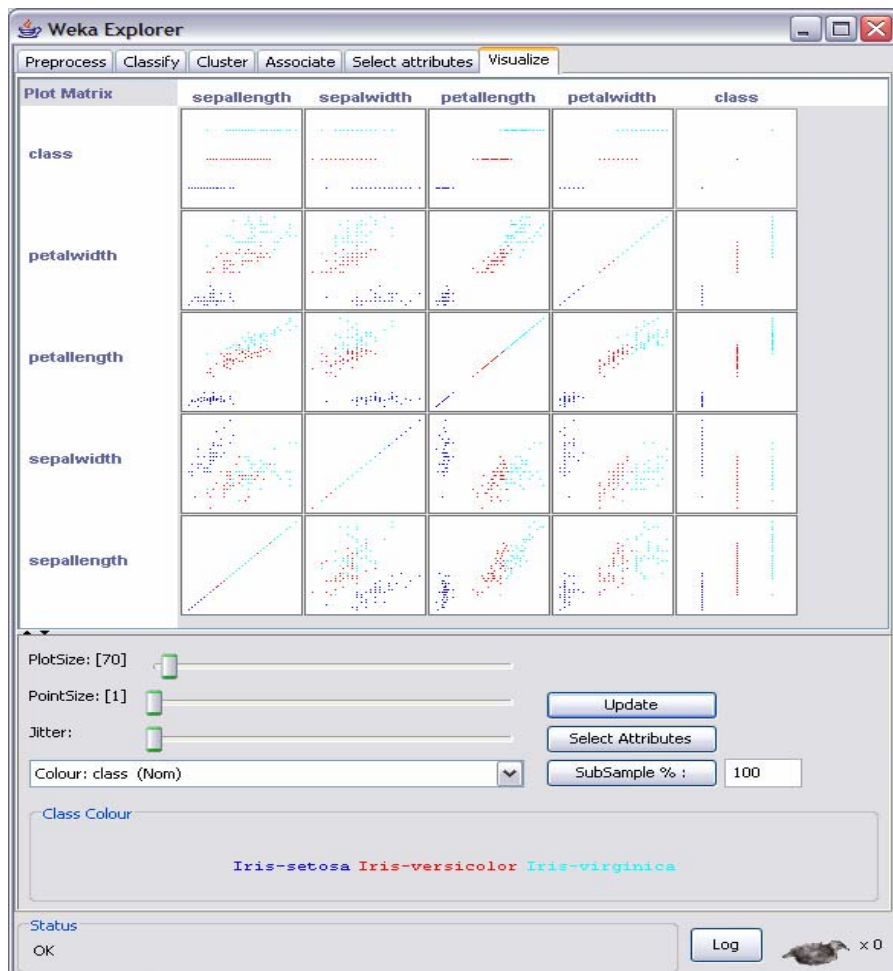
The screenshot shows the Weka Explorer interface with the 'Selected attribute' panel for 'sepalength'. The histogram below shows the distribution of values for 'sepalength' across different classes. The x-axis represents the value of 'sepalength' (ranging from 4.3 to 7.9), and the y-axis represents the count of instances. The histogram is divided into three segments: blue (left), red (middle), and cyan (right). The counts for each segment are: 16 (blue), 30 (red), 34 (cyan), 28 (red), 25 (cyan), 10 (red), and 7 (cyan).

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

- Edit / Save
- Current relation
- Attributes
 - Select / remove
 - All / none / invert
- Selected attribute
 - Name / type / missing / distinct / unique
 - Statistics
- Visualize class / all



Οπτικοποίηση



- Σκεδαστικά διαγράμματα για κάθε συνδυασμό χαρακτηριστικών προς ανίχνευση συσχετίσεων
- Plot & point size
- Select attributes & colour
- Subsample
- Κλικ σε διάγραμμα → εστίαση με περαιτέρω επιλογές



Εργασία



Διαχείριση Λογαριασμών Πελατών (Customer Account Management)

- Μία επιχείρηση διατηρεί στοιχεία για κάθε πελάτη της
- Με βάση τα στοιχεία αυτά αποτιμά κάθε πελάτη ως 'καλό' ή 'κακό' ('good' / 'bad')
- Για παράδειγμα, μία τράπεζα ενδιαφέρεται να εντοπίσει τους κατόχους πιστωτικών καρτών που ενδέχεται να χρεοκοπήσουν
- Σημαντικό κόστος για κάθε λάθος υπόδειξη
 - Αν 'κακός' προβλέπεται ως 'καλός': απώλεια αποπληρωμής λόγω χρεοκοπίας
 - Αν 'καλός' προβλέπεται ως 'κακός': κοστοβόρα περαιτέρω εξακρίβωση χαρακτηριστικών πελάτη ή και διακοπή συνεργασίας με αξιόπιστο πελάτη



Σύνολο δεδομένων

- Δεδομένα εκπαίδευσης (training set)
 - 2528 υποδείγματα
 - 39 χαρακτηριστικά (δυναμικά, αιέραια, πραγματικά) & ζητούμενο (record label, good / bad)
- Δεδομένα επαλήθευσης (quiz set)
 - 1265 υποδείγματα
 - 39 χαρακτηριστικά
- Δεδομένα εξέτασης (test set)
 - 1265 υποδείγματα
 - 39 χαρακτηριστικά
- Καμία περαιτέρω πληροφορία περί του συνόλου των δεδομένων δεν είναι γνωστή (το σενάριο περί πιστωτικών καρτών επιλέχθηκε για λόγους επεξήγησης)



Στόχος

- Ποιοι είναι οι ‘κακοί’ πελάτες (bad record labels) του test set;
- Εξόρυξη γνώσης από τα δεδομένα εκπαίδευσης
- Ανάδραση της μαθησιακής διαδικασίας μέσω των δεδομένων του quiz set
 - Μπορεί να υποβληθεί προς αξιολόγηση μεγάλος αριθμός διαφορετικών συνόλων record labels του quiz set καθ’ όλη τη διάρκεια διεξαγωγής της εργασίας
- Τελική εφαρμογή των εξαγόμενων δομικών περιγραφών στα δεδομένα του test set



Προς υποβολή

- Πρόβλεψη good / bad accounts του test set
- Έκθεση αναλυτικής περιγραφής και αιτιολόγησης της διαδικασίας που επιλέχθηκε
- Παρουσίαση της μεθοδολογίας στους συναδέλφους / 'ανταγωνιστές'

Η βαθμολόγηση αποτελεί συνάρτηση του βαθμού αξιοπιστίας της πρόβλεψης, της λογικής τεκμηρίωσης της επιλεγμένης διαδικασίας και της πληρότητας της έκθεσης και παρουσίασης.



Τέλος

Επόμενη διάλεξη:
Προεπεξεργασία & Επιλογή Δεδομένων